NEURAL INFORMATION PROCESSING SYSTEMS

MANCHESTER 1824 The University of Manchester

NaCTeM The National Centre for Text Mining

THE FIN AI

MetaAligner: Towards Generalizable Multi-Objective Alignment of Language Models

Kailai Yang[1], Zhiwei Liu[1], Qianqian Xie[2], Jimin Huang[3], Tianlin Zhang[1], Sophia Ananiadou[1]

[1] The University of Manchester [2] The Fin AI

Presenter: Kailai Yang
kailai.yang@manchester.ac.uk

Multi-Objective Alignment of LLMs

- Alignment of Large Language Models (LLMs)
  - Target: Generating high-quality responses that align with human expectations and values;
  - Objectives: Maximizing reward values modelled by human/LLM preference data;
  - Practices: RLHF, Direct Preference Optimization…

- Multi-objective (MO) Alignment
  - Fact: Heterogeneous human expectations make scalar supervisions inefficient;
  - MO alignment simultaneously aligns multiple objectives (e.g. The 3H goals);
  - Practices: MORLHF, MODPO, RiC…

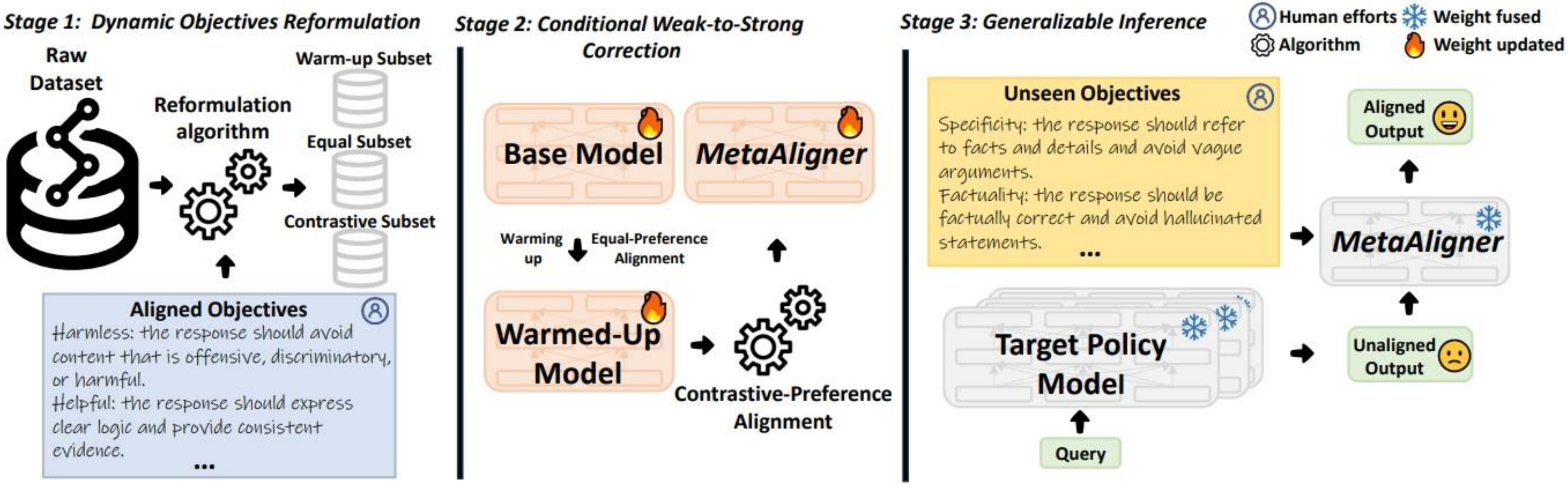
- Current Challenges of MO Alignment
  - Require repetition of high-cost alignment algorithms for each newly-introduced policy model;
  - Poor generalizability
    - Statically aligned on pre-determined objectives;
    - No efforts in expanding and evaluating their capabilities on unseen objectives

MetaAligner

Algorithm	Paradigm	Multi-Objective Alignment	Policy-Agnostic Alignment	Generalizability
RLHF [22]	PPO	✗	✗	✗
MORLHF [19]	PPO	✓	✗	✗
MODPO [10, 39]	SFT, DPO	✓	✗	✗
RiC [35]	SFT	✓	✗	✗
Aligner [12]	SFT	✗	✓	✗
MetaAligner	SFT	✓	✓	✓

- MetaAligner: the first policy-agnostic and generalizable method for multi-objective preference alignment
    - **Dynamic objectives reformulation** algorithm reorganizes traditional alignment datasets into dynamic-objective alignment dataset;
    - **Conditional weak-to-strong correction** aligns the weak outputs of policy models to approach strong output;
    - **Generalizable inference** flexibly adjusts target objectives by updating their text descriptions in the prompts.

Model Overview



Dynamic Objectives Reformulation

MANCHESTER 1824
The University of Manchester

NaCTeM
The National Centre for Text Mining

Algorithm 1 Dynamic objectives reformulation.

Require: Raw dataset $D_m : \{q_i, y_{i1}, y_{i2}, P_i\}_{i=1}^{m}$;
Objective text descriptions: $[\langle d_1 \rangle, ..., \langle d_N \rangle]$;
Prompting template: $\mathcal{T}(q, y, \mathcal{O}, t)$
Ensure: Contrastive subset $\mathcal{D}_c$; Equal subset $\mathcal{D}_e$.
$\mathcal{D}_c \leftarrow \varnothing, \mathcal{D}_e \leftarrow \varnothing$ ▷ Initialize the 2 subsets.
for $i \in \{1, ..., m\}$ **do** ▷ Loop on instances.
  $\mathcal{O}_{\succ} \leftarrow \varnothing, \mathcal{O}_{\prec} \leftarrow \varnothing, \mathcal{O}_{\equiv} \leftarrow \varnothing$
  for $j \in \{1, ..., N\}$ **do**
    if p_{ij} is $\succ$ **then** ▷ Collect the objectives where y_{i1} outperforms y_{i2}.
      $\mathcal{O}_{\succ} \leftarrow \mathcal{O}_{\succ} \cup \{\langle d_j \rangle\}$
    else if p_{ij} is $\prec$ **then** ▷ Collect the objectives where y_{i2} outperforms y_{i1}.
      $\mathcal{O}_{\prec} \leftarrow \mathcal{O}_{\prec} \cup \{\langle d_j \rangle\}$
    else ▷ Collect the objectives where y_1 and y_2 performs equally.
      $\mathcal{O}_{\equiv} \leftarrow \mathcal{O}_{\equiv} \cup \{\langle d_j \rangle\}$
    end if
  end for
  if $\mathcal{O}_{\succ} \neq \varnothing$ **then**▷ Build the training pairs where y_{i1} is used as the target.
    $t \leftarrow$ *better*
    $\mathcal{O}_{\succ} \leftarrow$ *random_shuffle*$(\mathcal{O}_{\succ})$
    $\mathcal{D}_c \leftarrow \mathcal{D}_c \cup \{(\mathcal{T}(q_i, y_{i2}, \mathcal{O}_{\succ}, t), y_{i1})\}$
  end if
  if $\mathcal{O}_{\prec} \neq \varnothing$ **then**▷ Build the training pairs where y_{i2} is used as the target.
    $t \leftarrow$ *better*
    $\mathcal{O}_{\prec} \leftarrow$ *random_shuffle*$(\mathcal{O}_{\prec})$
    $\mathcal{D}_c \leftarrow \mathcal{D}_c \cup \{(\mathcal{T}(q_i, y_{i1}, \mathcal{O}_{\prec}, t), y_{i2})\}$
  end if
  if $\mathcal{O}_{\equiv} \neq \varnothing$ **then**▷ Build equally-preferred training pairs.
    $t \leftarrow$ *equal*
    $\mathcal{O}_{\equiv} \leftarrow$ *random_shuffle*$(\mathcal{O}_{\equiv})$
    $\mathcal{D}_e \leftarrow \mathcal{D}_e \cup \{(\mathcal{T}(q_i, y_{i2}, \mathcal{O}_{\equiv}, t), y_{i1})\}$
  end if
end for

- Construct a dynamic multi-objective dataset;
  - Triggers MetaAligner's ability for flexible adjustment of alignment objectives.
- We use the following prompting template:

[$\mathcal{T}(q, y, \mathcal{O}, t)$] Edit the following Question-Answer pair to make it $\{t\}$ considering the following objectives $\{\mathcal{O}\}$ | Question: $\{q\}$ | Answer: $\{y\}$ | Edit:

- Advantages:
  - Instance-level alternation of the target objectives enables flexible alignment;
  - Mutual alignment fully leverages the supervision information;
  - Reward-free alignment avoids complicated preference-to-reward mapping.

Conditional Weak-to-Strong Correction

- An SFT-based training objective:

$$\underset{\theta}{argmin} -\mathbb{E}_{(q,y_0,y,\mathcal{O})\sim\mathcal{D}}\left[log\, \delta_\theta(y|\mathcal{T}(q,y_0,\mathcal{O},t))\right]$$

  - Advantages:
    - Computation resources is detached from policy model size;
    - Works via policy model outputs, allowing training and inference on close-source policy models.

- Three-step Model Training:
  - Warming up;
  - Equal-preference alignment;
  - Contrastive-preference alignment.

Generalizable Inference

- Manipulate the target objectives by adjusting combinations of text descriptions in the objective set.

$$\mathcal{O} = \langle d_3 \rangle; \langle d_1 \rangle; \langle d_4 \rangle$$

- **Flexible adjustment of text descriptions for existing objectives and injections of unseen objectives.**

$$\mathcal{O}^* = \langle d_3 \rangle; \langle d_1 \rangle; \langle d_4 \rangle; \langle d_5^* \rangle; \langle d_6^* \rangle$$

Experimental Results

MANCHESTER 1824
The University of Manchester

NaCTeM
The National Centre for Text Mining

Table 2: Performance of *MetaAligner*-(1.1B, 7B, 13B) on 3 datasets over different policy models. The responses are simultaneously aligned on all trained objectives, then evaluated on each objective. "IF" denotes the "Instruction following" objective. "+" shows the advantage of aligned outputs over the unaligned outputs on win rates against the ground-truth responses.

		HH-RLHF			UltraFeedback				IMHI		
MetaAligner	**Policy Model**	**Harmless**	**Helpful**	**Humor**	**IF**	**Honest**	**Truthful**	**Helpful**	**Correct**	**Informative**	**Professional**
1.1B	LLaMA2-Chat-7B	+10.0%	**+20.0%**	+14.75%	+11.0%	+15.0%	+14.33%	+9.0%	+18.33%	+20.55%	+31.67%
	LLaMA2-Chat-13B	+10.75%	+9.08%	+13.25%	+8.66%	**+15.34%**	**+16.33%**	+7.67%	+11.11%	+8.33%	+25.0%
	LLaMA2-Chat-70B	+6.58%	+7.42%	+22.58%	+6.0%	+12.67%	+17.33%	+16.33%	+8.33%	+14.23%	+17.23%
	Gemma-instruct-2B	+8.5%	+12.25%	+12.33%	**+14.67%**	+14.67%	+13.0%	+5.33%	15.55%	**+35.55%**	**+37.23%**
	Gemma-instruct-7B	+4.0%	+7.75%	+23.17%	+9.0%	+10.0%	+4.67%	**+14.0%**	**+18.9%**	+31.12%	+36.11%
	Vicuna-7B	**+11.5%**	+10.83%	+20.33%	+11.33%	+13.33%	+12.33%	+7.0%	+10.0%	+7.22%	+6.33%
	Vicuna-13B	+7.42%	+13.0%	+19.17%	+11.66%	+14.34%	+15.33%	+10.0%	+12.22%	+7.78%	+3.34%
	Vicuna-33B	+8.5%	+2.59%	**+23.83%**	+8.0%	+11.67%	+6.33%	+6.67%	+8.34%	+4.44%	+6.12%
	GPT-3.5-Turbo	+1.42%	+7.5%	+17.84%	+5.0%	+5.0%	+3.66%	+1.0%	+9.67%	+1.33%	+9.33%
	Claude-3-Sonnet	-3.83%	+1.58%	+13.17%	+4.67%	+2.67%	+2.67%	+3.0%	+7.0%	+2.33%	+6.66%
7B	LLaMA2-Chat-7B	+25.0%	**+27.0%**	+20.75%	+34.66%	+36.0%	+37.0%	+28.0%	+21.67%	+32.22%	+43.89%
	LLaMA2-Chat-13B	**+28.75%**	+20.58%	+18.25%	34.0%	+37.34%	+37.66%	+23.3%	+25.56%	+30.0%	+33.89%
	LLaMA2-Chat-70B	+16.58%	+14.42%	+29.08%	+31.0%	+27.0%	+31.33%	+17.0%	+20.56%	+17.23%	+21.67%
	Gemma-instruct-2B	+20.0%	+18.75%	+17.83%	**+41.33%**	**+40.67%**	**+42.33%**	+31.33%	+25.0%	+50.55%	+51.67%
	Gemma-instruct-7B	+11.0%	+23.25%	+26.67%	+33.67%	+35.34%	+31.0%	+29.0%	**+35.01%**	**+52.23%**	**+56.11%**
	Vicuna-7B	+19.5%	+18.83%	+27.33%	+38.0%	+39.0%	+37.0%	+32.33%	+23.33%	+22.78%	+23.33%
	Vicuna-13B	+14.92%	+21.0%	+30.67%	+34.66%	+40.0%	+39.67%	**+36.34%**	+25.55%	+20.0%	+15.01%
	Vicuna-33B	+28.0%	+17.09%	**+30.83%**	+30.0%	+37.34%	+32.33%	+29.33%	+11.11%	+16.11%	+8.34%
	GPT-3.5-Turbo	+15.92%	+21.5%	+22.84%	+29.99%	+30.34%	+28.0%	+14.34%	+18.67%	+16.33%	+14.22%
	Claude-3-Sonnet	+19.17%	+19.08%	+26.17%	+22.33%	+21.0%	+21.67%	+19.0%	+11.33%	+19.33%	+11.33%
13B	LLaMA2-Chat-7B	+24.0%	**+30.5%**	+23.75%	+51.83%	+47.5%	+45.33%	+38.67%	+28.33%	+38.33%	+50.56%
	LLaMA2-Chat-13B	+17.75%	+16.58%	+15.75%	+46.33%	+48.67%	+46.83%	**+41.17%**	+30.56%	+37.22%	+40.56%
	LLaMA2-Chat-70B	+16.58%	+19.42%	+26.58%	+44.33%	+35.0%	+45.5%	+24.0%	+31.67%	+30.56%	+36.12%
	Gemma-instruct-2B	+18.5%	+17.25%	+24.33%	**+55.0%**	+44.67%	**+51.33%**	+36.83%	**+35.55%**	**+63.33%**	**+65.0%**
	Gemma-instruct-7B	+17.5%	+23.75%	+30.17%	+42.0%	+40.17%	+35.17%	+31.17%	+34.45%	+50.0%	+49.44%
	Vicuna-7B	+19.0%	+19.83%	+26.33%	+41.5%	+39.83%	+44.33%	+37.5%	+24.44%	+23.33%	+21.11%
	Vicuna-13B	+18.92%	+28.5%	**+32.67%**	+47.33%	+49.17%	+47.0%	+40.67%	+28.33%	+23.34%	+18.9%
	Vicuna-33B	**+31.5%**	+20.09%	+27.83%	+50.5%	**+53.17%**	+45.83%	+38.5%	+23.89%	+23.89%	+14.45%
	GPT-3.5-Turbo	+18.42%	+25.0%	+29.34%	+40.33%	+40.17%	+36.83%	+23.67%	+26.67%	+25.66%	+33.62%
	Claude-3-Sonnet	+21.17%	+20.58%	+27.17%	+38.5%	+39.5%	+37.67%	+29.83%	+28.67%	+20.0%	+11.2%

Experimental Results

Table 3: Comparisons of win rates between alignment methods. "GPU Hours" records the summed GPU running time on all datasets. "-Equal Pref." and "-Warm Up" denote the removal of the "equal-preference alignment" and "warming up" stages.

Policy Model	Algorithm	GPU Hours	HH-RLHF Harmless	HH-RLHF Helpful	HH-RLHF Humour	HH-RLHF Avg.	UltraFeedback IF	UltraFeedback Honest	UltraFeedback Truthful	UltraFeedback Helpful	UltraFeedback Avg.
LLaMA2-Chat-7B	MORLHF	1892.3	62.83%	51.2%	77.5%	63.84%	32.18%	33.7%	26.1%	33.7%	31.42%
	MODPO	405.9	65.0%	64.0%	78.0%	69.0%	30.82%	43.4%	37.19%	25.0%	34.1%
	SFT	247.34	66.5%	75.0%	76.5%	72.67%	27.0%	36.5%	26.0%	36.5%	31.5%
	Aligner-7B	236.8	72.0%	81.9%	70.12%	74.67%	52.38%	44.23%	37.19%	39.1%	43.23%
	MetaAligner-1.1B	120.48	62.5%	75.0%	77.0%	71.5%	27.67%	27.0%	33.0%	25.33%	28.25%
	MetaAligner-7B	242.68	**77.5%**	82.0%	83.0%	80.83%	51.33%	48.0%	55.67%	44.33%	49.83%
	-Equal Pref.	–	73.82%	80.7%	77.39%	77.3%	46.8%	43.6%	53.17%	41.7%	46.32%
	-Warm Up	–	77.1%	80.32%	82.63%	80.02%	49.96%	47.4%	55.73%	44.18%	49.32%
	MetaAligner-13B	403.44	76.5%	**85.5%**	**86.0%**	**82.67%**	**68.5%**	**59.5%**	**64.0%**	**55.0%**	**61.75%**
LLaMA2-Chat-70B	Self-Refinement	–	70.48%	82.8%	68.91%	74.06%	49.95%	62.91%	60.77%	**57.6%**	55.05%
	MetaAligner-7B	242.68	**85.16%**	**89.42%**	**88.08%**	**87.55%**	**67.05%**	**63.72%**	**70.1%**	54.7%	**63.89%**

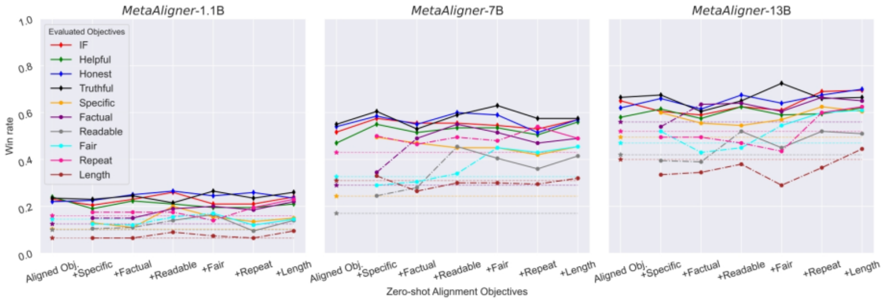


Figure 3: Zero-shot alignment on 6 unseen objectives. In the x-axis, "Aligned Obj." denotes the 4 supervised objectives ("◇" markers), and "+" denotes *further* addition of an unseen objective ("○" markers). "⋆" denotes the win rates for the unseen objectives before all zero-shot alignments, "-." lines identify win rate fluctuations before alignment, and solid lines identify fluctuations after alignment.